

BIROn - Birkbeck Institutional Research Online

Collins, E.J. and Brooms, Anthony C. (2005) The Bernoulli Feedback Queue with Balking: stochastic order results and equilibrium joining rules. Working Paper. Birkbeck, University of London, London, UK.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/26971/>

Usage Guidelines:

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>
contact lib-eprints@bbk.ac.uk.

or alternatively

ISSN 1745-8587



School of Economics, Mathematics and Statistics

BWPEF 0517

The Bernoulli Feedback Queue with Balking: Stochastic Order Results and Equilibrium Joining Rules

E.J. Collins
A.C. Brooms

November 2005

The Bernoulli Feedback Queue with Balking: Stochastic Order Results and Equilibrium Joining Rules

E. J. Collins* A. C. Brooms†

7 November, 2005

Abstract

We consider customer joining behaviour for a system that consists of a FCFS queue with Bernoulli feedback. A consequence of the feedback characteristic is that the sojourn time of a customer already in the system depends on the joining decisions taken by future arrivals to the system. By establishing stochastic order results for coupled versions of the system, we prove the existence, and uniqueness, of Nash equilibrium joining policies, and show that these are characterized by (possibly randomized) threshold rules. We contrast the Nash rule with the socially optimizing joining rule that minimizes the long-term expected average sojourn time (or cost) per customer. The latter rule is characterized by a non-randomized threshold, and we show that the Nash rule admits at least as many customers into the system as the socially optimizing one.

Keywords: FCFS queue with Bernoulli feedback; coupling; Nash equilibrium; social optimality

AMS: 90B22; 91A10; 60E15; 91A13; 91A14

1 Introduction

This paper considers the joining behaviour of customers into a First Come First Served Bernoulli Feedback queueing system. Each arriving customer joins the system, or balks, on the basis of the number of customers already present. It is assumed that customers who join the system do not renege at any stage. An important consequence of the Bernoulli feedback property is that the sojourn time of any customer who is already

*Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, U.K.

†School of Economics, Mathematics, & Statistics, Birkbeck College, Malet Street, London WC1E 7HX, U.K.

in the system may be affected by customer arrivals in the future. Joining behaviour of customers to the system is considered in the context of the following two scenarios. In the first, each customer compares their expected sojourn time (or cost) in the system with some fixed cost parameter associated with balking, and makes the joining decision that yields the smallest expected cost. Since this involves taking into account the joining decisions taken by other customers, it is natural to consider the Nash equilibrium as the appropriate characterization of behaviour. Our second scenario is one in which the joining decision of each customer is selected by a centralized authority, with the objective of minimizing long-term expected costs averaged across all customers. In common with other literature on admission control into queues, we discuss whether decentralized decision making can be as good as, or perhaps worse than, that under centralized control, when judged according to the social criterion posed in our problem.

Naor (1969) carried out one of the earliest studies of optimal customer joining behaviour into single-server queueing systems. He assumes a constant holding cost per customer per unit time and assumes that a fixed reward accrued to each customer in the system upon completion of service (thus, in effect, a linear holding cost). He shows that, within the class of (stationary) deterministic threshold policies, there exist unique individually optimal and socially optimal joining rules that minimize the expected cost to each customer and the long-run (expected) cost per unit time, respectively. Finally, he also shows that the socially optimal threshold is a lower bound on the threshold that is individually optimal.

Similar results have been established in a number of extensions to the above system. For example, Yechiali (1971) considers the GI/M/1 system (with linear cost structure), and shows that, amongst all policies, there exists a non-randomized threshold joining rule that is self-optimizing, from the point of view of each customer. He also shows that in the class of stationary policies, the socially optimizing policy that minimizes an average cost criterion, is also characterized by a non-randomized threshold. Again the socially optimal threshold is seen to be a lower bound to the one that is individually optimal. Yechiali (1972) establishes corresponding results for the GI/M/s queue. However, Altman & Hassin (2002) argue that the individually optimal joining policy for the $M/G/1$ queue does not exhibit the usual threshold structure, due to the queue lengths giving an indication as to the residual time of the customer in service to new arrivals at the system.

Using an approach based on uniformization (Lippman 1975), Lippman & Stidham (1977) derive results analogous to those of Naor and Yechiali for a model in which the service rate is a bounded, concave increasing, function of the number of customers in the system. Other relevant papers include Stidham (1978), where a convex holding cost is assumed, and Johansen & Stidham (1980), where a stochastic input-output system with a very general structure is considered. The survey article of Stidham (1985) and the book of Hassin & Haviv (2003) provide useful overviews of the relevant literature.

A common feature of all the models cited above is that the time or cost of a particular customer already in the system is unaffected by the joining behaviour of future

arrivals. This allows policies to be formulated that are optimal for each individual customer. However, in feedback models, the sojourn time of a customer already in the system depends on the joining decisions taken by future arrivals to the system. Natural applications of FCFS queueing systems with feedback arise, for example, in the theory of telephone traffic (Takács 1963); see Takagi (1991) and the references therein for variations and extensions to the basic model. We can also think of this system as a model for a single-line manufacturing process in which each job is independently tested, and sent through the process again if a fault is discovered or the work done to the job is deemed unsatisfactory (Peköz & Joglekar 2002). We can still define and construct 'optimal' joining rules for these models, but only if knowledge about the joining behaviour of future arrivals can be assumed; thus the appropriate solution concept to consider is that of the Nash equilibrium, and we discuss this in detail later in this paper.

Nash equilibrium joining rules for a 'single line' queueing system have been examined by Altman & Shimkin (1998) in the context of the processor sharing discipline. There it was assumed that the effective service rate to each customer in the queue, $\nu(x) = \mu(x)/x$, is strictly decreasing in x (where x is the number in the system, and $\mu(x)$ the corresponding service rate). For their system, they show that any candidate Nash equilibrium policy is characterized by a threshold structure, that a Nash equilibrium policy always exists, and will be unique when the policy is symmetric, i.e. each customer invokes exactly the same joining rule. This model was later extended to the case of multi-class heterogeneous preferences in Ben-Shahar, Orda & Shimkin (2000), in which the existence of the Nash equilibrium was also established.

The analysis of Altman & Shimkin (1998) can be modified and extended to deal with the multiple-server retrial queue (Brooms 2000), and the FCFS queue where the service rate is (strictly) decreasing in the number in the system (Brooms 2003). More recently, the FCFS queue with service rate strictly increasing in the queue length was analyzed in Brooms (2005). It was shown that, under the proviso that the joining rule for each customer is such that the chance that they are admitted to the queue is a non-increasing function of the queue length, there exists (at most) a finite number of symmetric Nash equilibria, and that at least one of these does not invoke randomization in its joining decisions. This should be contrasted with Altman & Shimkin (1998) in which existence and uniqueness were established (but with no guarantee of it being non-randomized) within the widest possible class of policies.

One of the difficulties in establishing the stochastic order relations required for our analysis stems from having to keep track of the actual position of certain of the customers in the system, due to the queue discipline; a similar difficulty is encountered for some other systems with the FCFS discipline (Brooms 2003, Brooms 2005), but not, for example, with processor sharing (Altman & Shimkin 1998), or retrial queues (Brooms 2000). Another difficulty stems from the Bernoulli feedback characteristic. A standard method for conducting sample path comparisons, it to generate coupled realizations of the queueing process; the progress of a 'marked' customer in each of the two processes is monitored and stochastic order results are thus derived. Unless considerable care is taken over the class of policies considered and over the type of

coupling used, 'dominance' across each and every pair of realizations is not achieved. So, in addition to the game-theoretic results presented in this paper, our second main contribution arises from the construction and use of apparently novel couplings in order to prove the ancillary lemmas.

The rest of this paper is organized as follows. In Section 2, the formulation of our model, a prescription of the joining rules to be used by customers, and a summary of the main results, are presented. In Section 3, sample path comparisons for our queueing process, and monotonicity results for the expected sojourn time in the system, as a function of the entry queue length x , are established. Similar results are proved with respect to the threshold value associated with symmetric threshold joining policies in Section 4; we also prove a continuity result for the expected sojourn time with respect to this threshold. We bring these results together to characterize the structure, and to prove the existence and uniqueness of a certain symmetric Nash equilibrium joining policy in Section 5. We also show that a certain socially optimizing policy can be characterized by a non-randomized threshold, and show that this is, in fact, a lower bound on the Nash threshold. We close the paper in Section 6, with some concluding remarks.

2 Preliminaries

2.1 The model

Consider a service system consisting of a single server queue (denoted by Q) with Bernoulli Feedback and First Come First Served (FCFS) queue discipline. Assume that each arriving customer joins the queue with a probability that depends only on the observed queue length x in Q just *prior* to their arrival at the system, and allow randomized decisions. A joining rule for an arriving customer is thus a sequence of numbers $\{u(x) \in [0, 1] : x = 0, 1, 2, \dots, B-1\}$, where B may be finite, or infinite; if the queue length just prior to their arrival is x then the customer joins the system with probability $u(x)$ and otherwise balks (i.e. does not join).

More formally, consider a process that starts at time $t = 0$ with an arriving customer C_0 that joins Q . We denote the subsequent arriving customers by C_1, C_2, \dots and let $X(t)$ denote the number of customers in Q at time t , with initial state $X(0) = x_0$.

Let $\mathcal{T} = \{T_1, T_2, T_3, \dots\}$ denote a sequence of independent, identically distributed, positive, continuous random variables, with finite expectation, which we interpret as the successive inter-arrival times, and let $\mathcal{W} = \{W_1, W_2, W_3, \dots\}$, and W , denote a sequence of independent, identically distributed, positive, continuous random variables, with finite expectation, which we interpret as the successive service times. The arrival epochs (to the system) of successive customers C_1, C_2, C_3, \dots are then given by the sequence $\mathcal{A} = \{A_1, A_2, A_3, \dots\}$, where $A_k = T_1 + \dots + T_k$, $k = 1, 2, \dots$

and, at least until the queue is empty for the first time, the successive service completion epochs in Q are given by the sequence $\mathcal{S} = \{S_0, S_1, S_2, S_3, \dots\}$, where $S_k = S_0 + W_1 + \dots + W_k$, $k = 1, 2, \dots$ (with appropriate modification thereafter). We assume that, with probability 1, the arrival epochs and service completion epochs are distinct.

Similarly, let $\mathcal{U} = \{U_1, U_2, U_3, \dots\}$ denote a sequence of independent random variables, each of which has a uniform distribution on the interval $(0, 1]$ and let $\mathcal{F} = \{F_0, F_1, F_2, \dots\}$ be a sequence of independent Bernoulli random variables with parameter p , so for each $k = 0, 1, 2, 3, \dots$, $F_k = 1$ with probability $p \in (0, 1)$ and $F_k = 0$ with probability $1 - p$. We interpret the U 's as the successive arrival joining decision variables, so customer C_k joins Q if and only if $U_k \leq u_k(X(A_k))$, and interpret the F 's as the successive feedback decision variables, so at the completion of the j -th service in Q after time $t = 0$, the customer that has just completed service is fed back to the end of the queue in Q if $F_j = 1$ and otherwise departs the system if $F_j = 0$.

In an abuse of terminology, we shall sometimes use Q to refer to the process as well as the queueing system itself; we shall refer to the number held in the system as the queue size or length (thus referring to the total number of customers queueing up for, and actually in, service). Under this model, the evolution of Q is completely determined by the initial queue size $X(0)$, the collection of joining rules for each one of the future customers $\{u_1, u_2, \dots\}$, the residual service time S_0 of the customer (if any) in service at Q at $t = 0$, and the values of the variables in the sequences $\mathcal{T}, \mathcal{W}, \mathcal{U}$ and \mathcal{F} . In particular, we assume $\{X(t) : t \geq 0\}$ is a left-continuous, piecewise constant process, whose jumps, if any, occur at arrival epochs $\{A_k\}$ or service completion epochs $\{S_j\}$, so that at S_j a customer is still with the server, whereas at S_j^+ the customer has either left the system or been fed back to the end of the queue. The jumps are formally described by the relations:

$$X(A_k^+) = X(A_k) + \mathbf{1}\{U_k \leq u_k(X(A_k))\} \quad k = 1, 2, 3, \dots \quad (1)$$

$$X(S_j^+) = X(S_j) - \mathbf{1}\{F_j = 0\} \quad j = 0, 1, 2, \dots \quad (2)$$

with appropriate modification if the buffer is full, or $X(S_j) = 0$, $j = 0, 1, 2, \dots$

2.2 Individual joining rules and population policies

Let u denote the joining rule for a given customer. We are particularly interested in the simple class of *threshold* joining rules under which a customer joins Q if the queue size is below a given threshold value, balks if the queue size is above the threshold value, and possibly randomizes between these actions if it equals the threshold value. Let \mathbb{Z}^+ denote the set of integers $\{1, 2, 3, \dots\}$ and let \mathbb{N} denote $\mathbb{Z}^+ \cup \{0\}$.

For nonnegative integer $L \in \mathbb{N}$ and $q \in [0, 1)$, we say a joining rule u is an $[L, q]$ -threshold rule if for $x \in \mathbb{N}$

$$u(x) = \begin{cases} 1 & \text{if } x < L \\ q & \text{if } x = L \\ 0 & \text{if } x > L \end{cases} \quad (3)$$

Associated with each $[L, q]$ -threshold rule is a unique real value $g = L + q$. We refer to g as the *threshold value* associated with the rule, and represent the rule itself more compactly by $[g]$.

For a population of customers arriving in the sequence C_0, C_1, C_2, \dots , we call the corresponding vector of customer joining rules a *population joining policy* and denote it by $\pi = (u_0, u_1, u_2, \dots)$. We let \mathbb{D}^∞ denote the class of *non-increasing* population policies for which each component rule u_k is such that $u_k(x)$ is non-increasing in x ; we let \mathbb{S}^∞ denote the class of *symmetric* population policies for which each of the components rules u_k are identical; and we let \mathbb{T}^∞ denote the class of *threshold* population policies, for which each u_k is a threshold joining rule. Observe that $\mathbb{T}^\infty \subset \mathbb{D}^\infty$. If all customers adopt the same joining rule u then we denote the resulting population joining policy $\pi = (u, u, u, \dots) \in \mathbb{S}^\infty$ by u^∞ ; similarly, if all customers use the same threshold joining rule $[g]$ we denote the resulting population joining policy by $\pi = [g]^\infty$.

2.3 Main Results

In the following sections, we prove a number of stochastic order results pertaining to the behaviour of the expected sojourn time of a customer in the system. Apart from being of interest in their own right, these results will be used to establish the existence, uniqueness, and structure of Nash equilibrium population joining policies for an associated stationary game.

Let $v_k(x, \beta, \pi)$, $x \in \mathbb{N}$, be the sojourn time of C_k in Q , given that at its arrival, x customers were already present in the system, the residual service time of the customer at the server is $\beta > 0$, and that future arrivals adhere to the decision rules inferred by π . Define $V_k(x, \beta, \pi)$ to be the expected value of $v_k(x, \beta, \pi)$. When the service time has an exponential distribution, the expected sojourn time of a customer that joins the queue does not depend on the residual service time (if any), and we simply write $v_k(x, \pi)$ and $V_k(x, \pi)$ respectively.

Note: indexing of entry queue sizes of the form $x \in \mathbb{N}$, $x = 0, 1, \dots$, or $x = 1, 2, \dots$ are to be understood as running up to $B - 1$ whenever B is finite. Also, the interval $[0, B)$ is interpreted to mean $[0, B]$ if B is finite, and $[0, \infty)$ if B is infinite.

Our main results are listed in the rest of this section. Theorems 1 to 4 characterize

the dependence of the expected sojourn time on both x and g , and are mostly proved by invoking couplings of a non-trivial nature. The game-theoretic results of Theorems 5 and 6 are proved using a combination of Theorems 1-4, but under the proviso that the total expected time spent at the server for a customer in Q is less than the 'cost' of balking from the system.

Theorem 1 *Consider a $GI/G/1$ Bernoulli feedback system and let $\pi \in \mathbb{D}^\infty$ be any non-increasing population joining policy. Then, for each $x = 1, 2, \dots$ and $\beta > 0$, $V(x+1, \beta, \pi) - V(x, \beta, \pi) \geq (1-p)E(W)$.*

The specialization of this result to the case of exponential service times can be found in Corollary 3.1. Theorem 1 is somewhat less general than its counterpart in Altman & Shimkin (1998) in that π is restricted to lie in \mathbb{D}^∞ . The class \mathbb{D}^∞ infers that there is less chance that each customer actually joins the system as the queue length there increases. Under additional assumptions on the arrival and departure processes, we can extend our result to another class of policies.

Theorem 2 *Consider an $M/M/1$ Bernoulli feedback system and let $\pi \in \mathbb{S}^\infty$ be any symmetric population joining policy. Then, for each $x = 0, 1, 2, \dots$, $V(x+1, \pi) - V(x, \pi) \geq (1-p)E(W)$.*

Theorem 3 *Consider a $GI/G/1$ Bernoulli feedback system and let $[g]^\infty$ and $[\tilde{g}]^\infty$ be symmetric threshold population joining policies with $0 \leq g < \tilde{g}$ and $\tilde{g} \in [0, B)$.*

- (i) *Suppose $\tilde{g} \leq 1$. Then $V(0, [\tilde{g}]^\infty) = V(0, [g]^\infty)$, and for each $x = 1, 2, \dots$ and $\beta > 0$, $V(x, \beta, [\tilde{g}]^\infty) = V(x, \beta, [g]^\infty)$.*
- (ii) *Suppose $g \geq 1$. Then there exists $\delta_0 > 0$ such that $V(0, [\tilde{g}]^\infty) - V(0, [g]^\infty) \geq \delta_0$, and for each $x = 1, 2, \dots$ and $\beta > 0$, there exists $\delta_x > 0$ such that $V(x, \beta, [\tilde{g}]^\infty) - V(x, \beta, [g]^\infty) \geq \delta_x$.*

Theorem 4 *Consider a $GI/G/1$ Bernoulli feedback system and let $[g]^\infty$ be a symmetric threshold population joining policy with $g > 0$. Then $V(0, [g]^\infty)$ is a continuous function of g , and, for each $x = 1, 2, \dots$ and $\beta > 0$, $V(x, \beta, [g]^\infty)$ is a continuous function of $g \in [0, B)$.*

Theorem 5 *Consider a $GI/M/1$ Bernoulli feedback system and assume that attention is restricted to the class \mathbb{D}^∞ of non-increasing population joining policies.*

- (i) *If $\pi = (u_0, u_1, u_2, \dots) \in \mathbb{D}^\infty$ is a Nash equilibrium population joining policy, then each u_k is a threshold joining rule (with finite threshold).*
- (ii) *There exists a unique symmetric Nash equilibrium population joining policy $\pi^* = (g^*, g^*, g^*, \dots) = [g^*]^\infty$ in the class of policies \mathbb{D}^∞ .*

Theorem 6 *Consider an $M/M/1$ Bernoulli feedback system.*

- (i) *If $\pi = u^\infty \in \mathbb{S}^\infty$ is a Nash equilibrium population joining policy, then u is a threshold joining rule (with finite threshold).*
- (ii) *There exists a unique symmetric Nash equilibrium population joining policy $\pi^* = (g^*, g^*, g^*, \dots) = [g^*]^\infty$.*

3 Monotonicity in the queue length x

3.1 Monotonicity for a GI/G/1 system

We first consider a $GI/G/1$ Bernoulli feedback queueing system where each potential customer uses a joining rule which is a non-increasing function of the queue size just prior to their arrival. Let x denote the queue size upon joining. We show that, for $x \geq 1$, the expected sojourn time of a joining customer is a strictly increasing function of x .

Without loss of generality, we focus on a marked customer C that joins the queue at time $t = 0$. For $k = 1, 2, 3, \dots$, we assume each successive potential customer, C_k , say, arrives at corresponding epoch A_k , and finds a queue of size $X(A_k)$. C_k has the option of either joining the queue or departing the system, and joins the queue with probability $u_k(X(A_k))$, where each $u_k(x)$ is a (possibly different) non-increasing function of x . Note that the presence of a finite buffer B can be incorporated by taking $u_k(x) = 0$ for $x \geq B$.

Let $v(x, \beta, \pi)$ denote the sojourn time for customer C who joins the queueing system, when the queue size just prior to arrival is $x \geq 1$, the population joining policy (i.e. the set of joining rules for later arriving potential customers) is $\pi = (u_1, u_2, u_3, \dots)$, and when the residual service time for the customer currently in service at time $t = 0$ is $S_0 = \beta > 0$. Let $V(x, \beta, \pi)$ be the expected value of this quantity.

To compare $v(x, \beta, \pi)$ with $v(x + 1, \beta, \pi)$, we look at path-wise comparisons of coupled realizations of two queueing processes, say Q and \tilde{Q} , in which marked customers C (resp. \tilde{C}) join the queue at time $t = 0$ when there are already x (resp. $x + 1$) customers in the queue, the population joining policy is π and the current residual service time is β . We say that at each time t a customer in the Q process is *level* with a customer in the \tilde{Q} process if both have the same position (first, second, third etc.) in their respective queues, and we say one customer is *ahead of* (resp. *behind*) the other if it has a position nearer (resp. further from) its server. We show that for each sample path in the coupled processes, the customer who joins with x in the system leaves either at the same epoch or at least one service completion before the customer who joins with $x + 1$ in the system. Moreover, this second possibility happens on a set of positive probability, so that $V(x, \beta, \pi) < V(x + 1, \beta, \pi)$.

The coupling we use here is designed to ensure that both C and \tilde{C} make the same number of visits to the server in the coupled systems. We saw from the model description in section 2.1 that the evolution of Q (and similarly \tilde{Q}) is completely specified by the sequence of successive inter-arrival times \mathcal{T} , service times \mathcal{W} , population joining policy π , joining decision random variables \mathcal{U} and feedback random variables \mathcal{F} . The coupling we use is defined in terms of these variables as follows:

Coupling 1 (i) Consider two processes Q and \tilde{Q} with $X(0) = x > 0$ and $\tilde{X}(0) = y > 0$. Couple the systems so that they have the same initial residual lifetime and so that, taken in the natural order, they have the same sequence of inter-arrival times, the same sequence of service times, the arriving customers use the same sequence of joining rules and the joining decision random variables take the same sequence of values. Formally, this means we set $S_0 = \tilde{S}_0 = \beta$, $T = \tilde{T}$, $\mathcal{W} = \tilde{\mathcal{W}}$, $\pi = \tilde{\pi}$ and $\mathcal{U} = \tilde{\mathcal{U}}$.

(ii) Now couple the feedback decision variables as follows. For $r = 1, 2, 3$ let $\mathcal{F}_r = \{F_{r,1}, F_{r,2}, F_{r,3} \dots\}$ denote mutually independent sequences of independent Bernoulli random variables, each with parameter p .

Use the sequence of values in \mathcal{F}_1 to determine both the successive feedback decisions for customer C in Q and the successive feedback decisions for \tilde{C} in \tilde{Q} , so, for example, both C and \tilde{C} are fed back after their first service if and only if $F_{1,1} \leq p$. Thus both C and \tilde{C} are fed back exactly the same number of times in both processes.

Use the sequence of values in \mathcal{F}_2 to determine the successive feedback decisions for all other customers in Q . Thus, the first customer in Q other than C to complete service is fed back if and only if $F_{2,1} \leq p$, the second is fed back if and only if $F_{2,2} \leq p$, etc.

Now consider the other customers in \tilde{Q} . By construction, the two processes Q and \tilde{Q} have the same service completion epochs, at least until one or other is empty for the first time. During this period, couple the feedback decision for each customer other than \tilde{C} to be exactly the same as that for the corresponding customer completing at the same time in Q , *except* for customers (other than \tilde{C}) who complete service at the same time as C . Say there is such a customer who completes service in \tilde{Q} at the same moment that C completes its k -th service in Q . Denote this customer by \tilde{H}_k and denote by H_k that customer in Q (if any) which is level with \tilde{C} at that moment. If such a customer H_k exists, define the feedback decision for \tilde{H}_k to be the same as the (already assigned) next feedback decision for H_k in Q . If there is no customer level with \tilde{C} at that moment, then define the feedback decision for \tilde{H}_k using the value of the k -th variable in the sequence \mathcal{F}_3 . Once the two processes no longer have the same service completion epochs, the feedback decisions can be assigned arbitrarily. \square

Note that under Coupling 1 a customer opposite C may depart even though C is fed back, so there may be epochs s when $X(s) > \tilde{X}(s)$. As well as showing how the relative positions of C and \tilde{C} are maintained between their service completion epochs, the next Lemma shows that if $\tilde{X}(0) = X(0) + 1$ then $X(s)$ can never exceed $\tilde{X}(s) + 1$.

Lemma 3.1 *Consider realizations of the two processes Q and \tilde{Q} under Coupling 1 with $y = x + 1$, and assume the population follows some non-increasing population policy $\pi \in \mathbb{D}^\infty$. Let τ denote the set of epochs at which C or \tilde{C} (or both) complete a service and neither have yet departed, and let s and t denote successive epochs in*

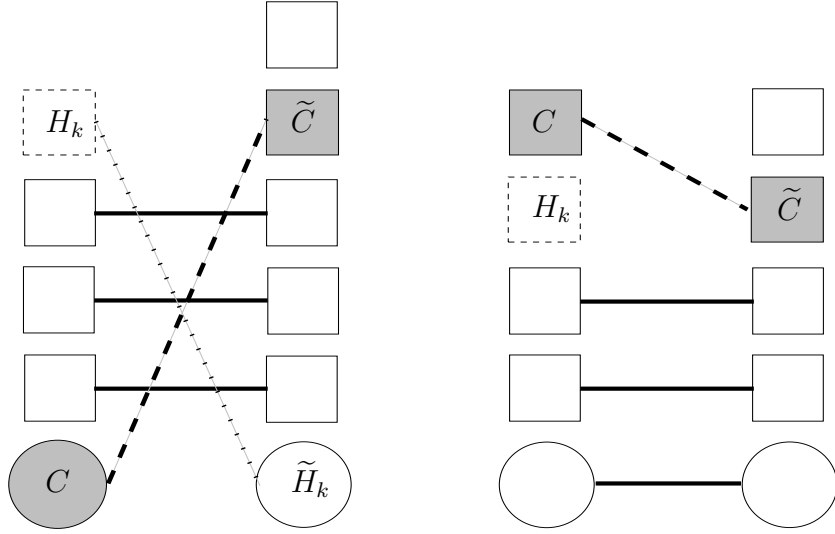


Figure 1: A possible realization of Q and \tilde{Q} just prior (L.H.S.) and just after (R.H.S.) C is fed back for the k -th time. C and \tilde{H}_k are in service on the L.H.S. The feedback decisions for C and \tilde{C} remain coupled throughout. The feedback decision for \tilde{H}_k is coupled with that of H_k if H_k is present, otherwise it is chosen independently; in the diagram neither are fed back. The next feedback decisions for the other customers in \tilde{Q} are coupled with those for the parallel customers in Q , and will be reassigned if they are fed back.

$\tau \cup \{0\}$. Then

- (i) The positions of C and \tilde{C} relative to each other do not change in (s, t) .
- (ii) If $\tilde{X}(s^+) \geq X(s^+)$ then $\tilde{X}(t) \geq X(t)$.
- (iii) If $X(s^+) = \tilde{X}(s^+) + 1$ then $X(t) \leq \tilde{X}(t) + 1$.

Proof

Consider the processes in the interval (s, t) , where any feedback decisions following the first service completion have been implemented by time s^+ , but those following the second service completion have not yet been implemented at t (by virtue of the 'left-continuity' of the queue-length process). During the interval, the composition of each queue changes only at arrival or service completion epochs.

(i): At service completion epochs, the coupling ensures that customers make the same feedback decision in both processes, so the positions of C and \tilde{C} relative to each other do not change. At arrival epochs, the arriving customers join behind C and \tilde{C} , so cannot affect their relative positions until the next epoch in τ . Thus the positions of C and \tilde{C} relative to each other do not change in (s, t) .

(ii) and (iii): At service completion epochs, the coupling ensures that the relative queue sizes remain unchanged. At arrival epochs when the queue lengths are equal, the coupling of the joining decision variables ensures that the same joining decision is

taken in both processes. At arrival epochs when one queue is smaller than the other, the fact that the joining decision rule is a non-increasing function of the size of the queue, together with the coupling and relation (1), ensures that either the same joining decision is taken in both processes or the arrival joins the queue in the process with the smaller queue but does not join in the process with the larger queue. Thus the difference in the queue sizes can only decrease during (s, t) and once the queue sizes are equal, they remain equal. In particular, if $X(s^+) = \tilde{X}(s^+) + 1$ then either $X(t) = \tilde{X}(t)$ or $X(t) = \tilde{X}(t) + 1$, so in either case $X(t) \leq \tilde{X}(t) + 1$. \square

Lemma 3.2 *Consider realizations of the two processes Q and \tilde{Q} under Coupling 1 with $y = x + 1$, and assume the population follows some non-increasing population policy $\pi \in \mathbb{D}^\infty$. Let K denote the common number of visits both C and \tilde{C} make to the server in each realization, and let s_1, \dots, s_K and $\tilde{s}_1, \dots, \tilde{s}_K$ denote the service completion epochs for C and \tilde{C} respectively. Then $\tilde{X}(s_k) \geq X(s_k)$ and $\tilde{s}_k \geq s_k$ for $k = 1, \dots, K$.*

Proof

For $k = 1, \dots, K$, let P_k denote the proposition: $\tilde{X}(s_k) \geq X(s_k)$ and $\tilde{s}_k \geq s_k$.

First assume $K = 1$. At $t = 0^+$, C has x other customers ahead of it in Q while \tilde{C} has $x + 1$ customers ahead of it in \tilde{Q} , so the position of C in Q is one ahead of that of \tilde{C} in \tilde{Q} . From Lemma 3.1, these relative positions are maintained until C completes service, so C leaves the system exactly one service completion epoch before \tilde{C} . Moreover, $\tilde{X}(0^+) = X(0^+) + 1 > X(0^+)$ so again from Lemma 3.1 $\tilde{X}(s_1) \geq X(s_1)$. Thus P_1 is true.

Now assume P_k is true for some $k = 1, \dots, K - 1$ for $K > 1$. Since $k < K$, both C and \tilde{C} are fed back after their k -th service. Now C is either level with \tilde{C} at s_k or C is ahead of \tilde{C} at s_k . If C is ahead of \tilde{C} at s_k , then there may or may not be a customer in Q level with \tilde{C} at s_k . If there is a customer in Q level with \tilde{C} at s_k , then that customer may or may not be fed back at its next service. There are then four cases to consider.

Case 1: [C is level with \tilde{C} at s_k].

Since C is level with \tilde{C} at s_k and $\tilde{s}_k \geq s_k$, both C and \tilde{C} are fed back together at s_k . Since $\tilde{X}(s_k) \geq X(s_k)$ and C was fed back with \tilde{C} at s_k , C is level with or ahead of \tilde{C} after being fed back, and $\tilde{X}(s_k^+) \geq X(s_k^+)$. Lemma 3.1 then implies that the next epoch in τ occurs at s_{k+1} , that C is still either level with or ahead of \tilde{C} at that point and that $\tilde{X}(s_{k+1}) \geq X(s_{k+1})$. Finally, \tilde{C} had completed no more than k services at s_k^+ so it must have completed no more than $k + 1$ services at s_{k+1}^+ , so $\tilde{s}_{k+1} \geq s_{k+1}$.

Case 2: [C is ahead of \tilde{C} at s_k and there is no customer in Q opposite \tilde{C} at s_k].

From the fact that there is no customer opposite \tilde{C} in Q when C completes service, it

follows immediately that: (i) $\tilde{X}(s_k) > X(s_k)$, (ii) C must be level with or ahead of \tilde{C} after being fed back, and (iii) the feedback decision for the customer \tilde{H}_k in \tilde{Q} who completes service at s_k is determined by the corresponding value in the sequence \mathcal{F}_3 , independent of the realization for Q . Since $\tilde{X}(s_k) > X(s_k)$, we have $\tilde{X}(s_k^+) \geq X(s_k^+)$ whether \tilde{H}_k departs or is fed back. Since C is level with or ahead of \tilde{C} at s_k^+ , Lemma 3.1 implies that the next epoch in τ is at s_{k+1} , that C is still level with or ahead of \tilde{C} at that point, and that $\tilde{X}(s_{k+1}) \geq X(s_{k+1})$. Since C was ahead of \tilde{C} at s_k and $\tilde{s}_k \geq s_k$, \tilde{C} must have completed at least one less service than C at s_k^+ , so it must still have completed at least one less service than C at s_{k+1}^+ , giving $\tilde{s}_{k+1} > s_{k+1}$.

Case 3: [C is ahead of \tilde{C} at s_k , H_k is opposite \tilde{C} at s_k and is fed back at its next service].

Since C is ahead of \tilde{C} at s_k then, together with $\tilde{s}_k > s_k$, this implies that \tilde{C} must have completed say $(r - 1)$ services at s_k^+ , where $(r - 1) < k$. Since C is ahead of \tilde{C} at s_k , there is a customer $\tilde{H}_k \neq \tilde{C}$ in \tilde{Q} who completes service at s_k and whose feedback decision is coupled to be the same as that for H_k , i.e. \tilde{H}_k is also fed back at s_k^+ . Thus $\tilde{X}(s_k^+) \geq X(s_k^+)$. Since there was a customer level with \tilde{C} at s_k , C is now behind \tilde{C} after the feedback. Lemma 3.1 then implies that the next epoch in τ occurs when \tilde{C} completes service at \tilde{s}_r and that $\tilde{X}(\tilde{s}_r) \geq X(\tilde{s}_r)$. At \tilde{s}_r^+ , \tilde{C} has completed $r \leq k < K$ services, so both \tilde{C} and H_k are fed back, giving $\tilde{X}(\tilde{s}_r^+) \geq X(\tilde{s}_r^+)$. Since $\tilde{X}(\tilde{s}_r) \geq X(\tilde{s}_r)$, C is now ahead of \tilde{C} after the feedback. Lemma 3.1 then implies that the next epoch in τ occurs when C completes service at s_{k+1} , that C is still ahead of \tilde{C} at that point, and that $\tilde{X}(s_{k+1}) \geq X(s_{k+1})$. Since \tilde{C} had completed $r \leq k$ services at \tilde{s}_r^+ and has not completed any more services by s_{k+1} , we have $\tilde{s}_{k+1} > s_{k+1}$.

Case 4: [C is ahead of \tilde{C} at s_k , H_k is opposite \tilde{C} at s_k and departs at its next service].

Since \tilde{H}_k now departs at s_k while C is fed back, we have $X(s_k^+) \leq \tilde{X}(s_k^+) + 1$ so either $X(s_k^+) \leq \tilde{X}(s_k^+)$ or $X(s_k^+) = \tilde{X}(s_k^+) + 1$. Since there was a customer level with \tilde{C} at s_k , C is now behind \tilde{C} after the feedback. Let r be as in Case 3. Lemma 3.1 now implies that the next epoch in τ occurs when \tilde{C} completes service at \tilde{s}_r , and that $X(\tilde{s}_r) \leq \tilde{X}(\tilde{s}_r) + 1$, so either $X(\tilde{s}_r) \leq \tilde{X}(\tilde{s}_r)$ or $X(\tilde{s}_r) = \tilde{X}(\tilde{s}_r) + 1$. At \tilde{s}_r^+ , \tilde{C} has completed $r \leq k < K$ services and so is fed back, while H_k departs just like \tilde{H}_k , so either $X(\tilde{s}_r^+) \leq \tilde{X}(\tilde{s}_r^+) - 1$ or $X(\tilde{s}_r^+) = \tilde{X}(\tilde{s}_r^+)$, i.e. $\tilde{X}(\tilde{s}_r^+) \geq X(\tilde{s}_r^+)$. Thus \tilde{C} is either fed back level with C or behind C . Lemma 3.1 now implies that the next epoch in τ is at s_{k+1} , that C is still level with or ahead of \tilde{C} at that point, and that $\tilde{X}(s_{k+1}) \geq X(s_{k+1})$. Since \tilde{C} had completed less than k services at s_k^+ and has only completed one service between s_k and s_{k+1} , it has completed at most $k + 1$ services by s_{k+1}^+ , and so $\tilde{s}_{k+1} \geq s_{k+1}$.

Thus in all cases P_k implies P_{k+1} . Since P_1 is true (using a similar argument for

establishing P_1 when $K = 1$), the result follows by induction. \square

Theorem 1

Consider a GI/G/1 Bernoulli feedback system and let $\pi \in \mathbb{D}^\infty$ be any non-increasing population joining policy. Then, for each $x = 1, 2, \dots$ and $\beta > 0$, $V(x + 1, \beta, \pi) - V(x, \beta, \pi) \geq (1 - p)E(W)$.

Proof

Consider realizations of the two processes Q and \tilde{Q} as in Coupling 1. Assume that there are initially x customers ahead of C in Q and $y = x + 1$ customers ahead of \tilde{C} in \tilde{Q} and that customers in both Q and \tilde{Q} are using the same non-increasing population joining policy $\pi \in \mathbb{D}^\infty$. From Lemma 3.2, C completes its first service at s_1 (one customer ahead of \tilde{C}), and completes all its remaining services either level with \tilde{C} or at least one customer ahead. The probability that C (and \tilde{C}) depart after just one service is $(1 - p)$, and the expected extra time \tilde{C} spends in \tilde{Q} in that case is $E(W)$. Thus, taking expectation over all possible realizations, we have $V(x + 1, \beta, \pi) - V(x, \beta, \pi) \geq (1 - p)E(W)$. \square

3.2 Monotonicity for a GI/M/1 system

When the service time has an exponential distribution, the residual service time of a customer in service at an arrival epoch has exactly the same exponential distribution as the service time of a customer starting service at that point. Thus the expected sojourn time of a customer that joins the queue does not depend on the residual service time of the customer (if any) in service on joining. In this case we can write $V(x, \pi)$ for the expected sojourn time for customer C when the queue size on joining is x and the population joining policy is $\pi = (u_1, u_2, u_3, \dots)$.

Corollary 3.1

Consider a GI/M/1 Bernoulli feedback system and let $\pi \in \mathbb{D}^\infty$ be any non-increasing population joining policy. Then, for each $x = 0, 1, 2, \dots$, $V(x + 1, \pi) - V(x, \pi) \geq (1 - p)E(W)$.

Proof

The proof for $x = 1, 2, \dots$ follows directly from Theorem 1 since the expected sojourn times are independent of β . Moreover, the result for $x = 0$ can be proved in exactly the same way as the results for $x > 0$ in section 3.1, since we can now arrange the coupling so that the residual service time of the customer in service in \tilde{Q} at $t = 0$ has exactly the same value as the service time of the customer joining and entering service in Q at $t = 0$. \square

3.3 Monotonicity for an M/M/1 system

When the arrival process forms a stationary Poisson process we can extend the class of population joining rules for which Theorem 1 applies. Consider an $M/M/1$ Bernoulli feedback system where potential customers all use the same joining rule u , where $u(x)$ is a general (not necessarily non-increasing) function of the queue size x on arrival. We again show that the expected sojourn time of a customer that joins a non-empty queue is a strictly increasing function of the queue size on joining.

Again let $v(x, \pi)$ denote the sojourn time for customer C when the queue size on joining is $x \geq 1$, when the symmetric population joining policy (for arriving potential customers) is $\pi = u^\infty$, and let $V(x, \pi)$ be the expectation of this quantity. Again we compare $v(x, \pi)$ with $v(x + 1, \pi)$, by looking at path-wise comparisons of coupled realizations of two queueing processes, say Q and \tilde{Q} , in which marked customers C (resp. \tilde{C}) join the queue at $t = 0$ when there are already x (resp. $x + 1$) customers in the queue.

The coupling we use is, perhaps, more complex than Coupling 1, but is again designed to ensure that both C and \tilde{C} make the same number of visits to the server in the coupled systems.

For fixed u , the evolution of \tilde{Q} is completely specified as before by the sequence of successive inter-arrival times \tilde{T} , service times \tilde{W} , joining decision random variables \tilde{U} and feedback random variables \tilde{F} . The coupled evolution of Q can then be described informally as follows: Consider a realization of \tilde{Q} in which \tilde{C} makes K visits to the server. For $k = 1, 2, 3, \dots$ let $\tilde{s}_1, \dots, \tilde{s}_K$ denote the corresponding service completion epochs of \tilde{C} . We "freeze" the process Q until \tilde{C} is level with C and then couple the two processes to have the same arrival epochs, service completion epochs, arrival decision variables and feedback decision variables until both C and \tilde{C} complete their first service. By construction, when \tilde{C} is fed back for the first time, there are at least as many customers ahead of it as there are ahead of C when it is fed back for the first time. To extend the realization until the next service completion epoch for C , again "freeze" the process Q until \tilde{C} is again level with C and then re-couple them until both C and \tilde{C} complete their next service. This procedure can be continued iteratively until both C and \tilde{C} depart.

We can define this coupling more formally as follows:

Coupling 2

Let s_1, \dots, s_K and $\tilde{s}_1, \dots, \tilde{s}_K$ be the successive service completion epochs of customers C and \tilde{C} , respectively, and set $s_0 = \tilde{s}_0 := 0$. For some $k \in \{0, \dots, K - 1\}$, assume that we have constructed Q up to the epoch s_k^+ , $\tilde{X}(\tilde{s}_k) \geq X(s_k)$ and $\tilde{s}_k \geq s_k$.

Set $b = \tilde{X}(\tilde{s}_k) - 1 - (X(s_k) - 1) = \tilde{X}(\tilde{s}_k) - X(s_k)$, which for $k \geq 1$ (resp. $k = 0$) rep-

resents the difference between the number ahead of C and the number ahead of \tilde{C} as they are fed back for the k -th time (resp. as they join their respective systems at time 0).

Now observe \tilde{Q} from \tilde{s}_k^+ until b services have taken place and then couple Q with it. Let r_1, r_2, \dots denote the arrival epochs of successive customers in \tilde{C} after \tilde{s}_k and t_1, t_2, \dots the successive service completion epochs. Assume that there have been e arrivals and f services in \tilde{Q} prior to \tilde{s}_k^+ , and that there are a arrivals and b service completions in \tilde{Q} in the interval $(\tilde{s}_k, t_b]$ and c arrivals and d service completions in the interval $(t_b, \tilde{s}_{k+1}]$, so $d = X(s_k^+)$ and $t_{b+d} = \tilde{s}_{k+1}$. Then starting at time s_k^+ , we construct the realization of Q over the interval $(s_k, s_k + t_{b+d} - t_b]$ as follows. If $c > 0$, then there are taken to be c arrivals in Q in this interval, with arrival epochs $s_k + r_{a+1} - t_b, \dots, s_k + r_{a+c} - t_b$ and joining decision parameters $U_{e+a+1}, \dots, U_{e+a+c}$. There are taken to be d service completions in Q in this interval, with service completion epochs $s_k + t_{b+1} - t_b, \dots, s_k + t_{b+d} - t_b$ and feedback decision parameters $F_{f+b+1}, \dots, F_{f+b+d}$.

The coupling after s_K is arbitrary.

□

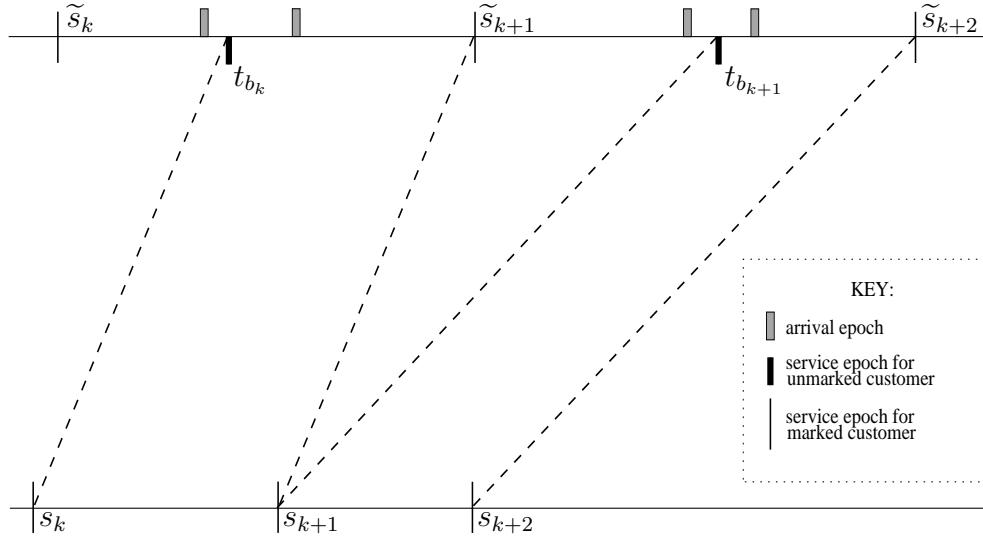


Figure 2: Possible realizations of Q (bottom) and \tilde{Q} (top) under Coupling 2. The diagram shows the time horizons near the k -th service transitions of the marked customer in the two processes. The service epochs for which \tilde{C} becomes level with C after the k -th and $(k + 1)$ -th services of \tilde{C} are given by t_{b_k} and $t_{b_{k+1}}$, respectively. The arrival epochs in \tilde{Q} closest to t_{b_k} and $t_{b_{k+1}}$ are also depicted.

Theorem 2

Consider an $M/M/1$ Bernoulli feedback system and let $\pi \in \mathbb{S}^\infty$ be any symmetric population joining policy. Then, for each $x = 0, 1, 2, \dots$, $V(x+1, \pi) - V(x, \pi) \geq (1-p)E(W)$.

Proof

Consider realizations of the two processes Q and \tilde{Q} under Coupling 2. Assume that there are initially x customers ahead of C in Q and $x+1$ customers ahead of \tilde{C} in \tilde{Q} and that all customers in both Q and \tilde{Q} use the decision rule inferred by the symmetric policy $\pi \in \mathbb{S}^\infty$.

For $k = 1, \dots, K$, let P_k denote the proposition: $\tilde{X}(\tilde{s}_k) \geq X(s_k)$ and $\tilde{s}_k \geq s_k$.

Assume that $K > 1$ and that P_k holds for some $k \in \{1, \dots, K-1\}$.

Due to the coupling, the position of C in Q at s_k^+ is exactly the same as that of \tilde{C} in \tilde{Q} at t_b^+ and their relative positions stay the same over the respective intervals $(s_k, s_{k+1}]$ and $(t_b, \tilde{s}_{k+1}]$. The last service completion in \tilde{Q} in the interval $(t_b, \tilde{s}_{k+1}]$ occurs when \tilde{C} completes its next service, so C completes its next service at the corresponding epoch and $s_{k+1} = s_k + t_{b+d} - t_b$. At that point C is either fed back in the same way as \tilde{C} if $k+1 < K$ or C departs like \tilde{C} if $k+1 = K$.

The arrival, service completion, and feedback processes, for Q over the interval $(s_k, s_k + t_{b+d} - t_b]$ completely mirror those in \tilde{Q} over the interval $(t_b, t_{b+d}]$. However, the number $\tilde{X}(t_b^+)$ in \tilde{Q} at t_b^+ is, by construction, at least as great as $X(s_k^+)$ in Q . Furthermore, consider any $t \in (0, t_{b+d} - t_b)$. Then while $\tilde{X}(t_b + t) > X(s_k + t)$, the actual queue size dependent joining decision in Q may differ from the corresponding decision in \tilde{Q} ; however, if for some $t^* \in (0, t_{b+d} - t_b)$ the queue sizes are the same (i.e. $\tilde{X}(t_b + t^*) = X(s_k + t^*)$), then the joining decisions will be the same for all $t \in [t^*, t_{b+d} - t_b]$, and hence the queue sizes will stay equal over the corresponding intervals in Q and \tilde{Q} . Thus, by construction, $\tilde{X}(\tilde{s}_{k+1}) \geq X(s_{k+1})$. Finally, by assumption, $\tilde{s}_k \geq s_k$ and by construction $t_b \geq \tilde{s}_k$, so that $\tilde{s}_{k+1} = t_{b+d} = t_b + (t_{b+d} - t_b) \geq s_k + (t_{b+d} - t_b) = s_{k+1}$. Thus P_{k+1} also holds.

By construction, $\tilde{X}(\tilde{s}_0) = \tilde{X}(0) = x+1 > x = X(0) = X(s_0)$, C starts $b = 1$ customer ahead of \tilde{C} in their respective systems, and completes its first service at $s_1 = \tilde{s}_1 - t_1$ where t_1 is the service completion epoch of the first customer served in \tilde{Q} after s_0^+ . Using a similar argument to the one in the preceding paragraph, it also follows that $\tilde{X}(\tilde{s}_1) \geq X(s_1)$. Thus P_1 holds here (and in the case where $K = 1$). Hence, and in particular, $\tilde{s}_K \geq s_K$.

The probability that C (and \tilde{C}) depart after just one service is $(1-p)$, and the ex-

pected extra time \tilde{C} spends in \tilde{Q} in that case is $E(W)$.

Now, for each k , the memoryless property of the Exponential distribution implies that the value $r_{a+1} - t_b$ used in constructing the arrival epochs for the interval (s_k, s_{k+1}) is again an independent observation from the same Exponential inter-arrival distribution. Thus, when we take expectation over all possible realizations of \tilde{Q} the coupling also generates an expectation over all possible realizations of Q with just the right distributions for the inter-arrival (and service) times. Thus $V(x+1, \pi) - V(x, \pi) \geq (1-p)E(W)$. \square

4 Monotonicity and continuity in the threshold g

In this section we again consider a $GI/G/1$ Bernoulli feedback queueing system but now assume all customers use the same threshold joining rule $[L, q]$. Recall from section 2.2 that the rule can be written in compact form as $[g]$, where $g = L + q$. We consider the dependence of the expected sojourn time on the joining rule and show that it is a continuous function of g , which is constant for $g \in [0, 1]$, and is strictly increasing for $g \geq 1$.

To motivate the population joining rule, consider what would happen if, instead of joining the feedback queue, customers could join an alternative queueing system where the expected sojourn time was fixed at θ . We assume customers always join the feedback system when it is empty on arrival. However, if the queue size on arrival is $x \geq 1$, we assume that each arriving customer joins the feedback queue only if their expected sojourn time is less than the fixed sojourn time in the alternative queue. In this case, the results of the previous section mean that each customer will use a threshold joining rule. Our focus here is on the behaviour of the expected sojourn time of an individual customer that does join the feedback queue when all the other customers are using the same threshold joining rule $[g]$.

Now let $g = L + q$ and $\tilde{g} = \tilde{L} + \tilde{q}$ denote the threshold values for two threshold joining rules with $g < \tilde{g}$, so that either $L < \tilde{L}$ or $L = \tilde{L}$ and $q < \tilde{q}$. Let $v(x, \beta, [g]^\infty)$ (resp. $v(x, \beta, [\tilde{g}]^\infty)$) denote the sojourn time for a customer who joins when there are already $x \geq 1$ customers in the system, when all other customers are using joining rule $[g]$ (resp. $[\tilde{g}]$) and the customer in service on joining has residual service time β . Let the expected value of $v(x, \beta, [g]^\infty)$ (resp. $v(x, \beta, [\tilde{g}]^\infty)$) be denoted by $V(x, \beta, [g]^\infty)$ (resp. $V(x, \beta, [\tilde{g}]^\infty)$).

To compare $v(x, \beta, [g]^\infty)$ and $v(x, \beta, [\tilde{g}]^\infty)$, we again compare coupled realizations of two processes. We show that in the coupled processes the customer who joins the system in which customers use $[g]$ leaves either at the same epoch or at least one service completion epoch before the customer who joins the system in which customers

use $[\tilde{g}]$. We then show that this second possibility happens on a set of positive probability, so that $V(x, \beta, [\tilde{g}]^\infty) > V(x, \beta, [g]^\infty)$.

Assume that there are initially x customers ahead of both C in Q and \tilde{C} in \tilde{Q} . Assume also that all other customers in Q use the same threshold joining policy $\pi = [g]^\infty$ and all other customers in \tilde{Q} use the same threshold joining policy $\pi = [\tilde{g}]^\infty$, where $\tilde{g} > g$.

Lemma 4.1 *Consider realizations of the two processes Q and \tilde{Q} under Coupling 1 with $y = x$. Let τ denote the set of epochs at which C or \tilde{C} (or both) complete a service and neither have yet departed, and let s and t denote successive epochs in $\tau \cup \{0\}$. Then*

- (i) *the positions of C and \tilde{C} relative to each other do not change in (s, t)*
- (ii) *if $\tilde{X}(s^+) \geq X(s^+)$ then $\tilde{X}(t) \geq X(t)$*
- (iii) *if $X(s^+) = \tilde{X}(s^+) + 1$ then $X(t) \leq \tilde{X}(t) + 1$.*

Proof

The argument is exactly the same as that for Lemma 3.1, except for the part relating to the changes in the respective queue sizes at arrival epochs.

Under the given policies a customer arriving in Q at z when the queue size is x joins if and only if either $x < L$ or $x = L$ and $U \leq q$, and a customer arriving in \tilde{Q} at z when the queue size is x joins if and only if either $x < \tilde{L}$ or $x = \tilde{L}$ and $U \leq \tilde{q}$, where either $L < \tilde{L}$, or $L = \tilde{L}$ and $q < \tilde{q}$.

If $X(z) < \tilde{X}(z)$, then $X(z^+) \leq \tilde{X}(z^+)$, whatever the respective joining decisions. If $X(z) = \tilde{X}(z)$, then the customer will join in Q if and only if either $X(z) < L$ or $X(z) = L$ and $U \leq q$. Since $X(z) = \tilde{X}(z)$ and either $L < \tilde{L}$ or $L = \tilde{L}$ and $q < \tilde{q}$, the customer joining in Q implies either $\tilde{X}(z) < \tilde{L}$ or $\tilde{X}(z) = \tilde{L}$ and $U \leq \tilde{q}$, so the customer must also join in \tilde{Q} . Thus, at each arrival epoch in (s^+, t) , $X(z) \leq \tilde{X}(z)$ implies $X(z^+) \leq \tilde{X}(z^+)$, giving (ii).

Similarly, if $X(z) = \tilde{X}(z) + 1$, then the customer will join in Q only if the customer in \tilde{Q} also joins, so the customers either join in both queues (giving $X(z^+) = \tilde{X}(z^+) + 1$), neither queue, or just in \tilde{Q} (giving $X(z^+) = \tilde{X}(z^+)$). Combined with the argument used to establish (ii), this gives (iii). \square

Lemma 4.2 *Consider realizations of the two processes Q and \tilde{Q} under Coupling 1 with $y = x$. Let K denote the common number of visits both C and \tilde{C} make to the server in each realization, and let s_1, \dots, s_K and $\tilde{s}_1, \dots, \tilde{s}_K$ denote the service completion epochs for C and \tilde{C} respectively. Then $\tilde{X}(s_k) \geq X(s_k)$ and $\tilde{s}_k \geq s_k$ for $k = 1, \dots, K$.*

Proof

For $k = 1, \dots, K$, let P_k denote the proposition: $\tilde{X}(s_k) \geq X(s_k)$ and $\tilde{s}_k \geq s_k$.

First assume $K = 1$. At $t = 0$, C and \tilde{C} are level with x customers ahead of them. From Lemma 4.1, these relative positions are maintained until C completes service, so $s_1 = \tilde{s}_1$. Moreover, $\tilde{X}(0^+) = X(0^+)$ so again from Lemma 4.1 (ii), $\tilde{X}(s_1) \geq X(s_1)$. Thus P_1 is true.

The proof for the case $K > 1$ follows in exactly the same way as in Lemma 3.2, except that we invoke Lemma 4.1 instead of Lemma 3.1. \square

Theorem 3 Consider a $GI/G/1$ Bernoulli feedback system and let $[g]^\infty$ and $[\tilde{g}]^\infty$ be symmetric threshold population joining policies with $0 \leq g < \tilde{g}$ and $\tilde{g} \in [0, B)$.

(i) Suppose $\tilde{g} \leq 1$. Then $V(0, [\tilde{g}]^\infty) = V(0, [g]^\infty)$, and for each $x = 1, 2, \dots$ and $\beta > 0$, $V(x, \beta, [\tilde{g}]^\infty) = V(x, \beta, [g]^\infty)$.

(ii) Suppose $g \geq 1$. Then there exists $\delta_0 > 0$ such that $V(0, [\tilde{g}]^\infty) - V(0, [g]^\infty) \geq \delta_0$, and for each $x = 1, 2, \dots$ and $\beta > 0$, there exists $\delta_x > 0$ such that $V(x, \beta, [\tilde{g}]^\infty) - V(x, \beta, [g]^\infty) \geq \delta_x$.

Proof

Consider realizations of the two processes Q and \tilde{Q} under Coupling 1. Assume that there are initially x customers ahead of both C in Q and \tilde{C} in \tilde{Q} . Assume also that all other customers in Q are using the same threshold population joining policy $\pi = [g]^\infty$ and all other customers in \tilde{Q} are using the same threshold joining policy $\pi = [\tilde{g}]^\infty$, where $\tilde{g} > g$.

First suppose that $0 \leq g < \tilde{g} \leq 1$. The sojourn times of the marked customers in the two processes will differ only if there is a disparity in the queue lengths during their stay in the systems. A customer is admitted into the queue of either process only if the queue is empty just prior to arrival. Clearly, however, the marked customer will have left by then, thus establishing (i).

Let s_1 be as defined in Lemma 4.2. Now suppose that $1 \leq g < \tilde{g}$, and let R_x denote the set of realizations for which $X(s_1) = L$ and $\tilde{X}(s_1) = L + 1$. If $L < \tilde{L}$, then R_x would include for example realizations in which no customers arrived during the service periods of the first x customers, all these x customers departed following service, L customers arrived during the (first) service period for C (and hence \tilde{C}), and $q < U_L < 1$. If $L = \tilde{L}$, then R_x would include for example realizations in which no customers arrived during the service periods of the first x customers, all these x customers departed following service, L customers arrived during the (first) service period for C (and hence \tilde{C}), and $q < U_L < \tilde{q}$. Thus R_x has positive probability. Note that the event R_x is independent of the number of visits K that C and \tilde{C} make to the server and that $P(K = 2) = p(1 - p)$.

For realizations in R_x with $K = 2$, C departs Q at s_2 one service period ahead of \tilde{C} and the expected extra time \tilde{C} spends in \tilde{Q} in that case is $E(W)$. From Lemma 4.2, in all other realizations C completes all its services either level with \tilde{C} or at least one service period ahead. Thus, taking expectation over all possible realizations, we have $V(x, \beta, [\tilde{g}]^\infty) - V(x, \beta, [g]^\infty) \geq p(1 - p)P(R_x)E(W) = \delta_x > 0$. \square

We now introduce a third coupling which we will use to show that the expected sojourn time $V(x, \beta, \pi)$ is continuous in g for symmetric threshold policies $[g]^\infty$. The coupling is designed to ensure that the queue length in \tilde{Q} is no less than that of Q .

Coupling 3 Set $S_0 = \tilde{S}_0 = \beta$, $\mathcal{T} = \tilde{\mathcal{T}}$, $\mathcal{W} = \tilde{\mathcal{W}}$, $\mathcal{U} = \tilde{\mathcal{U}}$, $\mathcal{F} = \tilde{\mathcal{F}}$. \square

Under Coupling 3, the successive arrival epochs A_k and \tilde{A}_k are the same in both systems; the successive service completion epochs S_k and \tilde{S}_k are the same, at least until one or other system is empty; and the successive feedback variables are the same. However, although the successive joining variables U_k and \tilde{U}_k are the same, the successive arrival joining decisions will not necessarily be the same. C_k joins Q if and only if $U_k \leq u_k(X(A_k))$, and similarly for \tilde{C}_k . Thus the arrival joining decisions may differ in cases when the queue sizes $X(A_k)$ and $\tilde{X}(A_k)$ differ, or when the queue sizes are the same but the actions specified by the decision rules u_k and \tilde{u}_k differ.

Now consider realizations of the processes in Q and \tilde{Q} under Coupling 3, with $g = L + q$ and $\tilde{g} = L + \tilde{q}$, such that $0 \leq q < \tilde{q} < 1$, such that $\tilde{g} \in [0, B)$. This means that service and arrival events are identical under both processes, except that at queue length L an arriving customer in \tilde{Q} has a probability $(\tilde{q} - q)$ of being accepted when the corresponding customer is rejected in Q . The strategy will be to construct an upper bound on $V(x, \beta, [\tilde{g}]^\infty) - V(x, \beta, [g]^\infty)$ which can also be shown to tend to 0 as $\tilde{g} - g$ tends to 0.

Theorem 4 Consider a GI/G/1 Bernoulli feedback system and let $[g]^\infty$ be a symmetric threshold population joining policy with $g \in [0, B)$. Then $V(0, [g]^\infty)$ is a continuous function of g , and, for each $x = 1, 2, \dots$ and $\beta > 0$, $V(x, \beta, [g]^\infty)$ is a continuous function of g .

Proof

Consider realizations of the two processes Q and \tilde{Q} under Coupling 3, and policies $[g]^\infty$ and $[\tilde{g}]^\infty$, respectively, where g and \tilde{g} are as defined in the paragraph preceding the statement of this theorem. Assume that there are initially x customers ahead of C and \tilde{C} in their respective systems. From Theorem 3 part (i), continuity holds trivially on the interval $[0, 1]$. Thus assume that $1 \leq g < \tilde{g}$. By the coupling, C and \tilde{C} complete their first service at the same epoch ($s_1 = \tilde{s}_1$). For $k = 1, 2, \dots$, let E_k denote the set of realizations for which C and \tilde{C} complete their first k services at the same epochs

(so $s_1 = \tilde{s}_1, \dots, s_k = \tilde{s}_k$) but complete their $(k+1)$ -st service at different epochs ($s_{k+1} \neq \tilde{s}_{k+1}$). Let E_0 denote the remaining set of realizations for which C and \tilde{C} complete all their services at the same epochs, so E_0, E_1, \dots form a partition of the set of all possible realizations.

Because the two systems start in identical initial states and are coupled to have the same sequence of inter-arrival and service times, a realization in E_k ($k \geq 1$) occurs only if C is fed back at least k times, C and \tilde{C} have exactly the same service completion epochs s_r , $r = 1, \dots, k$, and there is at least one arrival in the period (s_{k-1}, s_k) who joins the system in \tilde{Q} but not in Q ; i.e. this customer arrives when there are L in both systems and has a joining decision variable U with $q < U \leq \tilde{q}$.

Let E_k^1 denote the event that C is fed back at least k times and C and \tilde{C} have exactly the same first k service completion epochs s_r , $r = 1, \dots, k$. Let E_k^2 denote the event that there is at least one arrival in the period (s_{k-1}, s_k) who joins the system in \tilde{Q} but not in Q , and let D denote the difference in the sojourn times of C and \tilde{C} . Then $E_k \subset E_k^1 \cap E_k^2$ so $P(E_k) \leq P(E_k^2|E_k^1)P(E_k^1)$ and $E(D) = \sum_{k=1}^{\infty} E(D|E_k)P(E_k) \leq \sum_{k=1}^{\infty} E(D|E_k)P(E_k^2|E_k^1)P(E_k^1)$.

Given that E_k happens, any difference in the sojourn time is due only to the difference between their sojourn times from s_k onwards. Since there can be at most $L+1$ customers in each system, the expected time C spends in the system between each service completion epoch is at most $(L+1)E(W)$ and the expected number of passes through the system after s_k is $1/(1-p)$, so the expected sojourn time of C from s_k onwards is no greater than $(L+1)E(W)/(1-p)$. Arguing similarly for \tilde{C} , $E(D|E_k)$ is at most $2(L+1)E(W)/(1-p)$.

Also, E_k^1 occurs only if C is fed back at least k times, so $P(E_k^1) \leq p^k$.

Finally, we derive a bound on $P(E_k^2|E_k^1)$ as follows. Consider an arrival process that starts with an arrival at time $t = 0$. Let Z denote a random variable independent of the arrival process whose distribution is the same as that of the sum of $L+1$ independent service times, and let Y denote the number of arrivals in the closed interval $[0, Z]$. Clearly Y is almost surely finite (Feller 1966) so $\sum_{r=0}^{\infty} P(Y = r) = 1$.

Now consider a realization in E_k^1 , so C and \tilde{C} are both fed back together to the end of their respective queues at $s_{k-1} = \tilde{s}_{k-1}$ and have the same k -th service completion epoch $s_k = \tilde{s}_k$. Since the population joining rules are threshold rules with threshold values of the form $g = L + q$ and $\tilde{g} = L + \tilde{q}$, the total number in each queue will be at most $L+1$ and so the time $s_k - s_{k-1}$ until their next service completion will be no more than the sum of $L+1$ independent service times and so will be stochastically no greater than Z . Moreover, the first subsequent arrival will occur after s_{k-1} so the num-

ber of arrivals in $[s_{k-1}, s_k]$ will be stochastically smaller than the number of arrivals in the interval $[0, s_k - s_{k-1}]$ for an arrival process that starts with an arrival at $t = 0$, and this will in turn be stochastically no greater than Y . Thus if M denotes the number of arrivals to (both) Q and \tilde{Q} in $[s_{k-1}, s_k]$, then M is stochastically smaller than Y . Since $[1 - (\tilde{q} - q)]^Y$ is strictly decreasing in Y (by noting that $[1 - (\tilde{q} - q)] < 1$), $E([1 - (\tilde{q} - q)]^M) \geq E([1 - (\tilde{q} - q)]^Y)$.

Let U_1, U_2, \dots be a sequence of independent random variables each with a Uniform distribution on $(0, 1]$. Think of U_r as the joining variable of the r -th arrival after s_{k-1} . Now given E_k^1 occurs, E_k^2 fails to occur if all joining decisions are the same in both systems in the interval $[s_{k-1}, s_k]$, which will follow if U_r does not lie in the interval $(q, \tilde{q}]$ for the r -th arrival in the interval, $r \geq 1$, since $X(s_{k-1}^+) = \tilde{X}(\tilde{s}_{k-1}^+)$. Thus, using the fact that the U_r are independent of all other variables, we have that for a given q and \tilde{q} ,

$$\begin{aligned}
1 - P(E_k^2|E_k^1) &\geq P(M = 0) + \sum_{r=1}^{\infty} P(M = r, \bigcap_{j=1}^r \{U_j \notin (q, \tilde{q}]\}) \\
&= P(M = 0) + \sum_{r=1}^{\infty} P(M = r) P(\bigcap_{j=1}^r \{U_j \notin (q, \tilde{q}]\}) \\
&= P(M = 0) + \sum_{r=1}^{\infty} \{1 - (\tilde{q} - q)\}^r P(M = r) \\
&= \sum_{r=0}^{\infty} (1 - (\tilde{q} - q))^r P(M = r) \\
&= \sum_{r=0}^{\infty} (1 - (\tilde{g} - g))^r P(M = r) \\
&= E[(1 - (\tilde{g} - g))^M].
\end{aligned}$$

It follows that $P(E_k^2|E_k^1) \leq 1 - E[(1 - (\tilde{g} - g))^M] \leq 1 - E[(1 - (\tilde{g} - g))^Y]$. However, $|(1 - (\tilde{g} - g))^Y| \leq 1$ and $(1 - (\tilde{g} - g))^Y \rightarrow 1$ as $\tilde{g} - g \rightarrow 0$ almost surely (using the fact that Y is almost surely finite). Hence, by the dominated convergence theorem, $E[(1 - (\tilde{g} - g))^Y] \rightarrow 1$ as $\tilde{g} - g \rightarrow 0$, and thus $\mathbb{P}(E_k^2|E_k^1) \rightarrow 0$ also. Thus

$$\begin{aligned}
E(D) &= \sum_{k=1}^{\infty} E(D|E_k) P(E_k) \\
&\leq \sum_{k=1}^{\infty} E(D|E_k) P(E_k^2|E_k^1) P(E_k^1) \\
&\leq [1 - E[(1 - (\tilde{g} - g))^Y]] [2(L + 1)E(W)/(1 - p)] \sum_{k=1}^{\infty} p^k \\
&\rightarrow 0 \quad \text{as} \quad \tilde{g} - g \rightarrow 0.
\end{aligned}$$

□

5 Individual Nash equilibrium and social optimality

So far, we have looked at joining decisions for an isolated Bernoulli feedback queue. We now assume that the cost of balking upon arrival to Q is some constant value θ . We can think of θ as the time spent (or, alternatively, the cost of) using a 'private' or self-service system which is slower than Q , in the sense that θ is greater than the total expected time spent at the server for each customer in Q . More precisely, it will be assumed that $1/\mu(1-p) < \theta$; this condition says that it is always optimal for a customer to join Q if there are no customers in the system upon arrival, and there will be no further customers joining the system in the future. The joining decision depends only on the observed number of customers at Q on arrival. Customers who join Q are not permitted to renege during their sojourn, nor are those who balk permitted to join Q at a later stage.

We consider first what happens when customers make their own individual joining decisions and each customer is only interested in minimizing their own expected sojourn time, or cost. Due to the Bernoulli feedback characteristic, the sojourn time of a particular customer in Q may be affected by the number of customers in the queue during its sojourn, which in turn is affected by the decisions of subsequent arriving customers. This problem fits into a game theoretic framework. We derive the Nash equilibrium solution for the state dependent stationary game that arises and show that under this regime, the joining rule for each customer has a particular (possibly randomized) threshold form.

We then consider what happens when the joining decision for each customer is made by a central controller or *social optimizer*, whose goal is to minimize the overall expected cost per customer, averaged across customers admitted to Q and those that balk. In this case we show that there is a deterministic threshold rule which characterizes a *socially optimal* joining rule.

Finally, we show that the threshold for the symmetric Nash equilibrium joining rule is at least as large as the threshold for the socially optimal rule. The interpretation is that, when other customers use the Nash equilibrium joining rule, it is not to the advantage of any particular customer to change their joining rule, even though the Nash equilibrium joining rule produces greater congestion in Q and greater overall average sojourn times than the socially optimal rule.

5.1 Individual Nash equilibrium

For customers who join Q , the sojourn time is given by the time interval between arrival at, and departure from, the system. For a $GI/M/1$ Bernoulli feedback system the expected sojourn time depends only on the population joining policy and the queue size on joining. Consider a customer C_k who arrives to find x customers already in Q when the population joining policy is π . Let $V_k(x, \pi)$ denote the expected sojourn time for customer C_k if they decide to join the system when there are already x in the system and the population joining policy is π . The overall expected time/cost spent to customer C_k if it invokes the joining rule u_k is

$$u_k(x)V_k(x, \pi) + (1 - u_k(x))\theta.$$

Consider an arbitrary population joining policy $\pi = (u_0, u_1, u_2, \dots)$. Each customer wishes to minimize their own expected sojourn time, or cost, in the light of the actions of other customers. The expected cost customer C_k if they join Q when the queue size is x is $V_k(x, \pi)$ and their expected cost if they decide to balk is θ . Thus we follow Ben-Shahar et al. (2000) in defining a joining rule u_k to be a *best response* for customer C_k against the policy π if:

$$u_k(x) = \begin{cases} 1 & \text{if } V_k(x, \pi) < \theta \\ q_x & \text{if } V_k(x, \pi) = \theta \\ 0 & \text{if } V_k(x, \pi) > \theta \end{cases} \quad (4)$$

where $0 \leq q_x \leq 1$ is arbitrary. A population joining policy $\pi = (u_0, u_1, u_2, \dots)$ is said to be a *Nash equilibrium* if, for every $k \in \mathbb{N}$, u_k is a best response for C_k against π . Thus no customer can gain by changing their own joining rule while other customers continue to use the Nash equilibrium policy. More precisely, for arbitrary $k \in \mathbb{N}$, the overall cost to C_k cannot be further minimized by replacing u_k with another joining rule.

Theorem 5 *Consider a $GI/M/1$ Bernoulli feedback system and assume that attention is restricted to the class \mathbb{D}^∞ of non-increasing population joining policies.*

- (i) *If $\pi = (u_0, u_1, u_2, \dots) \in \mathbb{D}^\infty$ is a Nash equilibrium population joining policy, then each u_k is a threshold joining rule (with finite threshold).*
- (ii) *There exists a unique symmetric Nash equilibrium population joining policy $\pi^* = (g^*, g^*, g^*, \dots) = [g^*]^\infty$ in the class of policies \mathbb{D}^∞ .*

Proof

Let $\pi = (u_0, u_1, u_2, \dots)$ be a non-increasing population joining policy. From Corollary 3.1 we have that, for each $k \in \mathbb{N}$, $V_k(x, \pi)$ is a strictly increasing function of $x \in \mathbb{N}$, with $V_k(x, \pi) \rightarrow \infty$ as $x \rightarrow \infty$.

Further, Theorem 3 and Theorem 4 together imply that $V_k(x, [g]^\infty)$ is constant for

$g \in [0, 1]$, strictly increasing in $g \in [1, B)$, and continuous for $g \in [0, B)$, for each $k \in \mathbb{N}$ and $x \in \mathbb{N}$.

Without loss of generality, we focus attention on customer C_0 , and consider the point-to-set mapping

$$G^*(g) = \{g' \in [0, B] : [g'] \text{ is optimal for } C_0 \text{ against } [g]^\infty\}.$$

If it were the case that $V_0(n, [0]^\infty) = \theta$ for some $n \geq 0$, then $V_0(n, [g]^\infty) = \theta$ for any $g \in (0, 1]$ also, due to the constancy of $V_0(\cdot, [g]^\infty)$ in this region. However, this would imply that the graph of $G^*(\cdot)$ would include the set of points in the box with corners $(0, n)$, $(0, n+1)$, $(1, n+1)$, and $(1, n)$. Non-intersection of this box with the line of unit slope, with the possible exception of $(1, n)$, can be guaranteed provided that $V_0(0, [0]^\infty) < \theta$; however, this is equivalent to the condition that $1/\mu(1-p) < \theta$. The rest of the proof is similar to that of Theorem 1 from Altman & Shimkin (1998). \square

The results of Theorem 5 are somewhat less general than their counterpart in Altman & Shimkin (1998) in that π is restricted to lie in \mathbb{D}^∞ . The class \mathbb{D}^∞ infers that there is less chance that each customer enters Q as the queue length there increases (which perhaps is not unreasonable). Nevertheless, we find that we can extend our result, under additional assumptions on the arrival process, to the class of \mathbb{S}^∞ .

Theorem 6 *Consider an $M/M/1$ Bernoulli feedback system.*

- (i) *If $\pi = u^\infty \in \mathbb{S}^\infty$ is a Nash equilibrium population joining policy, then u is a threshold joining rule (with finite threshold).*
- (ii) *There exists a unique symmetric Nash equilibrium population joining policy $\pi^* = (g^*, g^*, g^*, \dots) = [g^*]^\infty$.*

The proof of Theorem 6 is exactly the same as Theorem 5, except that it invokes Theorem 2 rather than Corollary 3.1.

5.2 Social optimality

We shall now look at the behaviour of the system when the joining decision for each customer is made by a central controller or *social optimizer*, on the basis of the queue length at Q just prior to the arrival of the customer. Again, customers who are not permitted to join Q will instead experience a fixed cost of θ . The goal of the social optimizer is to minimize the overall expected cost/sojourn time per customer, averaged across customers who are permitted to join Q , and those that are refused entry.

Let $J(X(A_n), a_n)$ represent the expected sojourn time of the n -th customer to arrive at the service facility when there are $X(A_n)$ customers in Q just prior to its arrival, the social optimizer takes decision a_n , and where the decisions of future arrivals are governed by the policy π (for conciseness of notation, this will be understood to coincide

with the policy over which expectation is taken in the cost function below). Without loss of generality, actions could be defined so that $a_n = 1$ corresponds to the customer being admitted into Q , and $a_n = 0$ to it being refused entry. In cases where the decision at time A_n is randomized, we can set $a_n = 0$, since the costs under both alternatives are equal.

Defining

$$\phi_\pi(i) = \lim_{n \rightarrow \infty} \inf E_\pi \frac{[\sum_{i=0}^n J(X(A_n), a_n) | X(0) = i]}{n+1}$$

then the social optimizer looks for a policy π^* such that

$$\phi_{\pi^*}(i) = \min \phi_\pi(i) \quad \text{for all } i \in \mathbb{N}.$$

Theorem 7 (i) Suppose Q is a $GI/M/1$ Bernoulli feedback system, where the service time distribution at each visit to the server is exponential with mean $1/\mu$. Then there exists a non-randomized threshold control rule, say with threshold N_s , that is socially optimal in the class of all (stationary) joining rules.

(ii) If the inter-arrival times are also exponential, say with mean $1/\lambda$, and if $\rho = \lambda/\mu(1-p) < 1$, then N_s is the socially optimal threshold if and only if

$$\frac{[N_s(1-\rho) - \rho(1-\rho^{N_s})]}{(1-\rho)^2} \leq \frac{\mu(1-p)}{\theta} < \frac{[(N_s+1)(1-\rho) - \rho(1-\rho^{N_s+1})]}{(1-\rho)^2}. \quad (5)$$

Proof

In Q the service time distribution at each visit to the server is exponential with mean $1/\mu$, so the total service time distribution (which excludes the time waiting for service) for each joining customer is again exponential with mean $1/(1-p)\mu$. Under any given centrally imposed joining rule for Q , the queue length process in Q is equivalent to that for a $GI/M/1$ system (say \hat{Q}) with the same joining rule but where each customer takes all their service periods consecutively, where their service time distribution is exponential with mean $1/(1-p)\mu$. Thus the distribution of the queue length as seen by an arriving customer, the evolution of the joining decisions, and the expected sojourn time averaged over all customers that join the system, have equivalent behaviour for Q and \hat{Q} (even though, for each x , the expected sojourn times for customers that join when there are x customers in the system Q will differ from the corresponding quantities for \hat{Q}). Hence the overall sojourn time, or cost, – averaged across customers who are admitted to the system and those that are refused entry – is the same for Q and \hat{Q} , and so the socially optimal joining rule is the same for both models.

The existence of a non-randomized socially optimal *threshold* control rule then follows from Theorem 6 of Yechiali (1971), where the quantity W_n used in relation (15) of that paper corresponds to $-J(X(A_n), a_n)$ here. For future reference, let us denote the corresponding threshold by N_s . When the inter-arrival times are also exponential,

say with mean $1/\lambda$, equation (5) characterizing the actual value of N_s can be established in a similar way to relation (22) in Naor (1969). \square

5.3 Comparison of Nash equilibrium and socially optimal policies

The Nash equilibrium population joining policy and the socially optimal joining policy described above are both threshold policies. In this section we show that the threshold N_s used by the socially optimal policy is no greater than the threshold g^* used by the Nash equilibrium policy.

If the social optimizer admits or rejects customers to Q according to the threshold control rule characterized by N_s , then this is exactly the same as the customers voluntarily adhering to the symmetric threshold joining policy $[N_s]^\infty$.

The proof of the following Lemma is based on the observation that under an appropriate coupling, if N_s were greater than g^* , then the queue length process associated with the first threshold would be greater than or equal to that of the second threshold, and showing that this leads to a contradiction.

Theorem 8

Assume $g^* \in [0, B)$.

- (i) Suppose Q corresponds to the $GI/M/1$ Bernoulli feedback system. Then $N_s \leq g^*$, where $[g^*]^\infty$ is the unique symmetric Nash equilibrium joining policy in the class \mathbb{D}^∞ .
- (ii) Suppose Q corresponds to the $M/M/1$ Bernoulli feedback system. Then $N_s \leq g^*$, where $[g^*]^\infty$ is the unique Nash equilibrium joining policy in the class \mathbb{S}^∞ .

Proof of Theorem 8

We define 2 processes:

the N_s -process:- where all customers use the policy $[N_s]^\infty$, and

the g^* -process:- where all customers use the policy $[g^*]^\infty$,

with the queue lengths initially equal to each other in the two processes.

Suppose for contradiction that $N_s > g^*$, where $N_s \in [0, B)$. Consider these two processes under Coupling 3.

Denote quantities associated with the socially optimal policy by an 's' and those by the Nash equilibrium with a '*'.

It is easy to show that

$$X^*(t) \leq X^s(t) \text{ for all } t \in [0, \infty). \quad (6)$$

As A_n is almost surely finite, $J(X(A_n), \cdot)$ is well-defined for each $n \geq 0$. Since corresponding customers (i.e. those with the same subscript index) across the two processes arrive at the service facility at the same time, it is sufficient to establish a dominance relation between the expected costs for each customer between these

processes. For simplicity, we will denote $J(X(A_n), a_n)$ by J_n . The following cases exhaust all possibilities for the n -th arrival to the system, $n \in \mathbb{N}$, i.e. customer C_n :

(a) Customer admitted into Q under both processes. Then

$$J_n^* = V_n(X^*(A_n), [g^*]^\infty) \leq V_n(X^s(A_n), [g^*]^\infty) < V_n(X^s(A_n), [N_s]^\infty) = J_n^s.$$

where the first inequality follows from (6).

(b) Customer rejected under both processes.

Then clearly

$$J_n^* = J_n^s = \theta.$$

(c) Customer admitted into Q under the g^* -process but rejected under the N_s -process.

$$J_n^* = V_n(X^*(A_n), [g^*]^\infty) \leq \theta = J_n^s$$

where the inequality follows from the fact that $[g^*]$ is the best response against $[g^*]^\infty$.

(d) Customer admitted into Q under the N_s -process but rejected under the g^* -process.

$$J_n^* = \theta \leq V_n(X^*(A_n), [g^*]^\infty) \leq V_n(X^s(A_n), [g^*]^\infty) < V_n(X^s(A_n), [N_s]^\infty) = J_n^s.$$

The first inequality follows from the fact that $[g^*]$ is the best response against $[g^*]^\infty$, the second from (6) and the monotonicity of $V_n(x, \cdot)$, and the third from the monotonicity of $V_n(\cdot, [g]^\infty)$.

Now consider the sequence of states $(X^*(A_n), X^s(A_n))$, $n \geq 0$, embedded at the arrival epochs. Clearly, this is a Markov Chain with a finite state space and with a single positive recurrent set $\mathcal{Z} = \{(i, j) : 0 \leq j - i \leq N_s - L'; i \leq L'\}$, where $L' = L^* + 1\{q^* > 0\}$. Furthermore, the states in \mathcal{Z} are aperiodic (since, for example, the state $(0, 0)$ is a member of \mathcal{Z} , and is aperiodic); therefore, for n sufficiently large, there exists an $0 < \varepsilon < 1$ such that the event $D_n = \{(X^*(A_n), X^s(A_n)) = (0, 0)\}$ occurs with a probability of at least ε ; thus case (a) occurs with at least probability ε for sufficiently large n . However, the inequalities of case (a), in conjunction with Theorem 3, can be used to show that $J_n^s - J_n^* \geq \delta > 0$, where $\delta = \inf\{\delta_x : x = 0, 1, \dots, N_s - 1\}$ (noting that $g^* \geq 1$). Hence, upon taking total expectations of J_n^s and J_n^* , we see that the socially optimal policy performs strictly worse than the Nash equilibrium threshold, giving the required contradiction.

□

6 Concluding Remarks

The analysis of this paper shows that, within the sub-class of symmetric policies that are characterized by a non-increasing joining rule, $\mathbb{D}^\infty \cap \mathbb{S}^\infty$, there exists a unique Nash equilibrium for the $GI/M/1$ Bernoulli feedback system. We also show that within the entire class of symmetric policies, \mathbb{S}^∞ , but under the additional assumption

of exponentiality for the inter-arrival times, there exists a unique Nash equilibrium. Under both of these regimes, the Nash equilibrium is characterized by a (possibly randomized) threshold joining rule. By a utilization of known results for the GI/M/1 queue, we establish (i) the existence and uniqueness of a joining rule, to be used by each customer, that minimizes the long-term expected average cost per customer, (ii) that the rule is characterized by a non-randomized threshold, and (iii) that a Nash equilibrium will admit a customer into the system whenever the socially optimal one does.

It is unclear, at this stage, however, whether the monotonicity results of Section 3 hold outside the class of policies so far considered. No counter-example is available at present to suggest that they do not hold outside the class.

7 Acknowledgements

The second author wishes to acknowledge the financial support of the Nuffield Foundation under Grant no. NAL/00721/G.

References

- Altman, E. & Hassin, R. (2002), ‘Non-threshold equilibrium for customers joining an M/G/1 queue’, *Proceedings of the 10th International Symposium of Dynamic Games, St. Petersburg, Russia*.
- Altman, E. & Shimkin, N. (1998), ‘Individual Equilibrium and Learning in Processor Sharing Systems’, *Operations Research* **46**(6), 776–784.
- Ben-Shahar, I., Orda, A. & Shimkin, N. (2000), ‘Dynamic Service Sharing with Heterogenous Preferences’, *Queueing Systems* **35**, 83–103.
- Brooms, A. C. (2000), ‘Individual Equilibrium Dynamic Routing in a Multiple Server Retrial Queue’, *Probability in the Engineering and Informational Sciences* **14**, 9–26.
- Brooms, A. C. (2003), ‘Assessing the Performance of a Shared Resource: Simulation vs. the Nash Equilibrium’, *YOR13, Keynote Proceedings* pp. 3–19.
- Brooms, A. C. (2005), ‘On the Nash equilibria for the FCFS queueing system with load-increasing service rate’, *Advances in Applied Probability* **37**, 461–481.
- Feller, W. (1966), *An Introduction to Probability Theory and its Applications*, Vol. 2, first edn, John Wiley and Sons, Inc., New York.
- Hassin, R. & Haviv, M. (2003), *To queue or not to queue: Equilibrium behavior in queueing systems*, Kluwer Academic Publishers.

- Johansen, S. G. & Stidham, S. (1980), 'Control of arrivals to a stochastic input-output system', *Advances in Applied Probability* **12**, 972–999.
- Lipmann, S. A. & Stidham, S. (1977), 'Individual versus Social Optimisation in Exponential Congestion Systems', *Operations Research* **25**, 233–247.
- Lippman, S. A. (1975), 'Applying a New Device in the Optimization of Exponential Queuing Systems', *Operations Research* **23**, 687–709.
- Naor, P. (1969), 'The Regulation of Queue Size by Levying Tolls', *Econometrica* **37**, 15–24.
- Peköz, E. & Joglekar, N. (2002), 'Poisson traffic flow in a general feedback queue', *Journal of Applied Probability* **39**, 630–636.
- Stidham, S. (1978), 'Socially and Individually Optimal Control of Arrivals to a GI/M/1 queue', *Management Science* **24**, 1598–1610.
- Stidham, S. (1985), 'Optimal Control of Admission to a Queueing System', *IEEE Trans. Auto. Control* **AC-30**, 705–713.
- Takács, L. (1963), 'A Single-Server Queue with Feedback', *Bell Systems Technical Journal* **42**, 505–519.
- Takagi, H. (1991), *Queueing Analysis: A Foundation of Performance Evaluation, Vol 1: Vacation and Priority Systems, Part 1*, Elsevier Science, Amsterdam.
- Yechiali, U. (1971), 'On Optimal Balking Rules for the GI/M/1 Queueing Process', *Operations Research* **19**, 349–370.
- Yechiali, U. (1972), 'Customers' optimal joining rules for the GI/M/s queue', *Management Science* **18**, 434–443.